# How to calculate an ANOVA table

## Calculations by Hand

We look at the following example: Let us say we measure the height of some plants under the effect of 3 different fertilizers.

| Treatment | Measures | | | Mean | $\hat{A}_i$ |
|---|---|---|---|---|---|
| X | 1 | 2 | 2 | ... | ... |
| Y | 5 | 6 | 5 | ... | ... |
| Z | 2 | 1 | | ... | ... |
| Overall mean | | // ... | | | |

**STEP 0:** The model:

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \tag{0.1}$$

$$\sum_i n_i A_i = 0 \tag{0.2}$$

Interpretation:
An observation $y_{ij}$ is given by: the average height of the plants ($\mu$), plus the effect of the fertilizer ($A_i$). and an "error" term ($\epsilon_{ij}$), i.e. every seed is different and therefore any plant will be different.
All these values ($\mu, A_i, \epsilon_{ij}$) are UNKNOWN!
Our GOAL is to test if the hypothesis $A_1 = A_2 = A_3 = 0$ is plausible[1].

**Remark 1** *If we have a control group (for example treatment "X" is "without any fertilizer", then we assume that the values of X are in some way the best approximation for $\mu$, therefore we can choose $A_1 = 0$ is spite of condition* (0.2).

**STEP 1:** complete the first table.
For the ***treatment means*** it is enough to calculate the mean of the values

$$\begin{aligned} Mean_X &= \frac{1+2+2}{3} = 1.667 \\ Mean_Y &= \frac{5+6+5}{3} = 5.333 \\ Mean_Z &= \frac{1+2}{2} = 1.5 \end{aligned}$$

[1] We DO NOT find "the correct value" for the $A_i$
We WILL NOT find *which* factor (treatment) has an effect, we just look if in general treatments has effect on the results.

The (estimated) overall mean ($\hat{\mu}$, which is an estimation of the exact, unknown overall mean $\mu$) is calculated as follows[2]:

$$\hat{\mu} = \frac{1+2+2+5+6+5+2+1}{8} = 3$$

The **estimated effects** $\hat{A}_i$ are the difference between the "estimated treatment mean" and the "estimated overall mean", i.e.

$$\hat{A}_i = Mean_i - \hat{\mu}$$

So

$$\hat{A}_1 = 1.667 - 3 = -1.333$$
$$\hat{A}_2 = 5.333 - 3 = 2.333$$
$$\hat{A}_3 = 1.5 - 3 = -1.5$$

Then:

| Treatment | Measures | | | Mean | $\hat{A}_i$ |
|---|---|---|---|---|---|
| X | 1 | 2 | 2 | 1.667 | -1.333 |
| Y | 5 | 6 | 5 | 5.333 | 2.333 |
| Z | 2 | 1 | | 1.5 | -1.5 |
| Overall mean | | | // | 3 | |

**STEP 2:** The ANOVA table.

| Cause of the variation | df | SS | MS | F | $F^{Krit}$ |
|---|---|---|---|---|---|
| Treatment | ... | ... | ... | ... | ... |
| Residuals | ... | ... | ... | | |
| Total | ... | ... | | | |

For the ***column** df* **(*degrees of freedom*)** just remember the rule "minus one":

$$\text{We have 3 different Treatments} \Rightarrow df_{treat} = 3 - 1 = 2$$
$$\text{We have 8 different measurements} \Rightarrow df_{tot} = 8 - 1 = 7$$
$$df_{treat} + df_{res} = df_{tot} \Rightarrow df_{res} = 7 - 2 = 5$$

For the ***column** SS* **(*sum of squares*)** we can proceed as follows:

[2] Remark that the overall mean does not necessary coincide with the mean of the $y_{i.}$!

$$\begin{aligned}SS_{treat} &= \text{"sum of squares \textbf{between} treatment groups"} \\ &= \sum \hat{A}_i^2 \cdot \#measures \\ &= (-1.33)^2 \cdot 3 + (2.33)^2 \cdot 3 + (1.5)^2 \cdot 2 = 26.17\end{aligned}$$

$$\begin{aligned}SS_{res} &= \text{"sum of squares \textbf{within} treatment groups"} \\ &= \sum_i \sum_j (y_{ij} - y_{i.})^2 = \sum_i SS_{row_i} \\ &= \left[(1-1.667)^2 + (2-1.667)^2 + (2-1.667)^2\right] + [0.667] + [0.5] \\ &= 1.83\end{aligned}$$

$$\begin{aligned}SS_{tot} &= \text{"\textbf{Total} sum of squares"} \\ &= \sum_{i,j} (y_{ij} - \hat{\mu})^2 \\ &= (1-3)^2 + (2-3)^2 + \ldots + (1-3)^2 = 28\end{aligned}$$

**Remark 2** *The total "SS" is always equal to the sum of the other "SS"!*

$$\begin{aligned}SS_{tot} &= SS_{treat} + SS_{res} \\ 28 &= 26.17 + 1.83\end{aligned}$$

For the ***column*** $MS$ **(*mean square*)** just remember the rule $MS = SS/df$, then:

$$\begin{aligned}MS_{treat} = \frac{SS_{treat}}{df_{treat}} &= \frac{26.17}{2} = 13.08 \\ MS_{res} = \frac{SS_{res}}{df_{res}} &= \frac{1.83}{5} = 0.37\end{aligned}$$

The ***F-value*** is just given by:

$$F = \frac{MS_{treat}}{MS_{res}} = \frac{13.08}{0.37} = 35.68$$

<u>Interpretation:</u>
The $F-value$ says us how far away we are from the hypothesis "we can not distinguish between error and treatment", i.e. "Treatment is not relevant according to our data"!
A big $F-$value implies that the effect of the treatment is relevant!

**Remark 3** *A small $F-$value does NOT imply that the hypothesis $A_i = 0\ \forall i$ is true. (We just can not conclude that it is false!)*

**STEP 3:** The decision:

Similar as for a T-test we calculate the critical value for the level $\alpha = 5\%$ with degrees of freedom 2 and 5 (just read off the values from the appropriate table)[3].

$$\alpha = 5\% \quad \Rightarrow \quad F_{2,5}^{krit}(5\%) = 5.79$$

We have calculated $F = 35.68 > F_{2,5}^{krit}(5\%)$.
Consequently we REJECT THE HYPOTHESIS $A_1 = A_2 = A_3 = 0$!!!
Similarly we could obtain the same result by calculating the $p - value$

$$p = 0.11\% \quad \Leftarrow \quad F_{2,5}(p) = 35.68$$

0.11% is less than 5%.
Consequently we reject the hypothesis $A_1 = A_2 = A_3 = 0$!!!

## Calculations with R

**STEP 0:** Insert the data

```
v <- c(1,2,2,5,6,5,2,1)
TR <- c(1,1,1,2,2,2,3,3)
d <- data.frame(v,TR)
d$TR <- as.factor(d$TR)
```

Interpretation:

- All the measurements have to be in the same vector (`v` in this case).
- For every factor (in this case just `TR`) we construct a vector, which can be interpreted as follows: the first three Values of the vector `v` belong to treatment 1 (X), the two last components to treatment 3 (Z) and the other 3 to treatment 2 (Y).
- WE know that `v` and `TR` belong to the same set of data, WE have to tell this even the PC! Therefore: `d <- data.frame(v,TR)`!
- WE know that the factor `TR` in the data set `d` is a factor, the PC doesn't! Therefore: `d$TR <- as.factor(d$TR)`!
- check with `str(d)` that `d$v` is a vector of numbers (`num`) and `d$TR` is a factor (`Factor`)

[3]Because $F$ is obtained by $MS_{treat}$ (2 deg of freedom) and $MS_{res}$ (5 deg of freedom), we calculate $F_{2,5}^{krit}(5\%)$.

```
> str(d)
'data.frame': 8 obs. of  2 variables:
 $ v : num  1 2 2 5 6 5 2 1
 $ TR: Factor w/ 3 levels "1","2","3": 1 1 1 2 2 2 3 3
```

**STEP 1:** Do the ANOVA table

```
d.fit <- aov(v~TR,data=d)
summary(d.fit)
```

Interpretation:

- Makes an ANOVA table of the data set `d`, analysing if the factor `TR` has a significant effect on `v`.
- The function `summary` shows the ANOVA table.

```
> summary(d.fit)
            Df  Sum Sq Mean Sq F value   Pr(>F)
TR           2 26.1667 13.0833  35.682 0.001097 **
Residuals    5  1.8333  0.3667
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

**STEP 2:** Decision:

Interpretation:

- Exactly the same as for the "by hand" calculated table
- With R we do not have the critical values to a level, but we have the $P-$value (`PR(>F)`).
  `PR(>F)`=0.1097%, this means: if we choose a level $a$ of 0.1%, we can not reject the Null-Hypothesis, by choosing a level $\alpha = 0.11\%$ or bigger we have to reject $H_0$! (Usually we choose $a = 5\% \Rightarrow H_0$ will be rejected!)